Statistics and Application 统计学与应用, 2020, 9(3), 491-497
Published Online June 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.93052

Hans汉斯

Hot Topics and Frontier Evolution of Text Mining in China

—A Visual Analysis of the Documents Collected by CNKI

Xin Wang[1], Jiangming Shen[2], Jiangnan Xu[1], Zhiyong Zeng[3*]

[1]Yunnan University Data Operation and Management Engineering Research Center, School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
[2]Enterprise Information Department of China Telecom Corporation Limited Yunnan Branch, Kunming Yunnan
[3]Yunnan University Data Operation and Management Engineering Research Center, School of Information, Yunnan University of Finance and Economics, Kunming Yunnan
Email: 364122330@qq.com, *zengzhiyong725@163.com

Received: Jun. 4th, 2020; accepted: Jun. 17th, 2020; published: Jun. 24th, 2020

Abstract

Data visualization software CiteSpace was used to analyze and study the institutions, authors and keywords of Chinese literature on text mining collected by CNKI. The research shows three conclusions: 1) There is little cooperation between research institutions. 2) There is little communication and cooperation among scholars, and the sense of cooperation still needs to be improved. 3) The hot topics include web mining, text classification, Chinese patent medicine, western medicine, data stratification algorithm, big data text and emotion analysis. Text mining and emotion analysis under big data are the main research trend of text mining in China.

Keywords

Text Mining, Visual Analysis, Hot Topic, Trend

国内文本挖掘的热点主题和前沿演进

——基于CNKI收录文献的可视化分析

王　鑫[1]，沈江明[2]，徐江南[1]，曾志勇[3*]

[1]云南财经大学统计与数学学院，云南省高校数据化运营管理工程研究中心，云南　昆明
[2]中国电信股份有限公司云南分公司企业信息化部，云南　昆明

*通讯作者。

文章引用：王鑫，沈江明，徐江南，曾志勇．国内文本挖掘的热点主题和前沿演进[J]．统计学与应用，2020，9(3)：491-497. DOI: 10.12677/sa.2020.93052

[3]云南财经大学信息学院，云南省高校数据化运营管理工程研究中心，云南 昆明
Email: 364122330@qq.com, *zengzhiyong725@163.com

收稿日期：2020年6月4日；录用日期：2020年6月17日；发布日期：2020年6月24日

摘 要

使用数据可视化软件CiteSpace基于中国学术网络出版总库(CNKI)收录的关于研究文本挖掘的中文文献对机构、作者、关键词等绘制图谱并进行分析与评述。经研究表现出三方面结论：1) 各研究机构之间合作比较分散，合作较少；2) 各学者间的交流与合作不显著，合作意识仍然有待提高；3) 研究的热点主题有web挖掘、文本分类、中成药、西药、数据分层算法、大数据文本、情感分析；大数据下文本挖掘与情感分析为我国文本挖掘研究的主要研究趋势。

关键词

文本挖掘，可视化分析，热点主题，趋势

1. 引言

近年来，随着计算机、大数据挖掘等技术的飞速发展，日常活动产出文本类电子日志与日俱增，使得获取大量电子文本成本变低。而各个领域文本日志、文本数据都或多或少蕴含着有利于该领域发展的潜藏信息，因此挖掘出其中蕴藏的信息财富变得尤为重要。这使得文本挖掘技术近年来在各个领域有极其显著的发展。而同时文本类数据相较于结构化数据的特殊性，特别是中文类文本，由于中文语义及中文文段没有像英文的天然空格作为分词等特点。对其进行分析往往需要更复杂的工序。因此文本挖掘技术广受计算机技术领域与数据挖掘领域的关注。为了迅速把握我国文本挖掘技术研究的热点主题及趋势。本文使用可视化软件 CiteSpace 对 CNKI 平台收录的有关文本挖掘技术的文献进行可视化定量分析[1]，期望能够为文本挖掘技术的后续研究提供趋势参考，这对了解目前文本挖掘前沿方向具有重要意义。

2. 研究工具及数据来源

(一) 研究工具。本文的数据处理采用的是由美国德雷塞尔大学的信息可视化专家陈超美教授开发的 CiteSpaceV 软件，该软件是在 JAVA 应用程序基础上开发的，能够对科学文献进行可视化分析，跟踪领域的研究热点，探测领域的研究趋势[2]。其特点是能把一个知识领域的大量文献数据以一种多元、分时、动态的可视化图形将该领域的发展历程集中展现在一张知识图谱上[3]。而本文是基于 CiteSpace5.6.R4 版本进行数据可视化，该版本容纳的数据量较之前版本大幅提升并加入支持图形旋转、压缩和舒展的功能[4]。从而为本文提供的文献作者、机构、关键字等数据，绘制更优的可视化图谱、建立节点间的关联来分析其之间的共现关系与共被引关系[5]。

(二) 数据来源。本文的数据全部来源于中国学术期刊网络出版总库(CNKI)收录的文献。使用 CNKI 的高级检索功能选择“文献”下的主题检索。在检索条件中填入“文本挖掘技术”“文本挖掘”的关键

字，为契合研究主题检索时间跨度选择的是 1998 年 4 月 1 日至 2020 年 4 月 18 日，且仅选择中文文献再经过人工筛选剔除新闻宣传等非本文研究重点的文献，最终保留有分析价值的文献 1168 篇。最后将目标文献以 Refworks 的参考文献格式导出并使用 CiteSpace 软件转码为其可以识别使用的文献格式。从而为后续的分析提供了数据支持。

3. 数据结果分析

(一) 发文量情况统计

1) 发文量与年份的统计分析。如图 1 显示了我国文本挖掘技术研究性文献在中国学术网络出版社总库(CNKI)发文量随年份的变化趋势。图中可以看到第一篇研究文本挖掘的文献是在 1999 年发布的。此后从 2001 年开始年发文量逐年快速线性上升且 2008 年突破了年发文量 80 篇，研究吸引力趋势可见一斑；而 2008 年到 2015 年的年度发文量保持在 60 篇到 90 篇的区间内。2016 年的发文量达到峰值 118 篇。2017 年开始发文量有所下跌但至今每年发文量依然保持在 90 篇以上的水平。由此可见文本挖掘技术受到了我国学者们的广泛关注与研究。

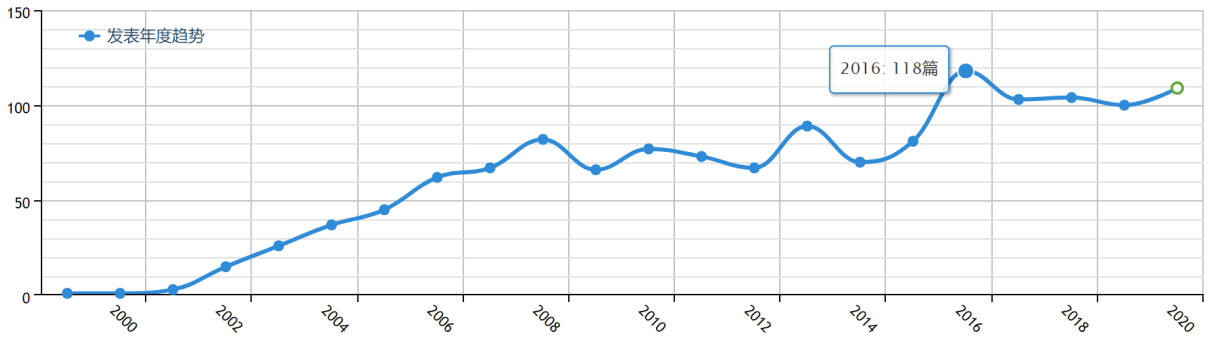

Figure 1. Trend of publication volume
图 1. 发文量年度趋势图

2) 机构群与作者群分析。在运行 CiteSpace 前，设置时间跨度为 1998 年至 2020 年，时间切片(Years Per Slice)为 1 年，节点类型为机构(Institution)，TOPN = 50，将阈值(Thresholding)的 c 设置为 2，2，20；cc 与 ccv 设置为 4，3，20。数据修剪(Pruning)勾选 Pathfinder 和 pruning sliced networks。其他参数均为默认值。经过运行得到如图 2 的机构共现可视化图谱。可以看到图谱中中国中医科学院中医临床基础医学研究所、兰州大学信息学院、上海中医药大学三所机构节点较大，由 CiteSpace 绘制的图片发文量越多节点越大的特点，可得这三所机构在我国文本挖掘的研究中发文量靠前、学术研究投入力度较大。此外，可以从该图谱中看到其节点数(N)为 324，连线数(E)为 175，密度(Density)为 0.0033。而节点间的连接线数表示节点即机构之间的联系，连接线越多表示机构间联系越紧密，图 2 中的连接数明显较少。其次，密度值越小节点间的联系越稀疏，从而也佐证了各机构间合作比较稀疏，合作较少，缺乏合作。

再次运行 CiteSpace，其他参数不变节点类型更改为作者(Author)。经过运行得到如图 3 的作者共现可视化图谱，从图中不难看出文献产出数量较多的是吕爱平、郭洪涛、正光和姜淼，其他作者如谭勇、杨静、吕诚、张弛等发文量次之。此外图中可以显著观察到其中图结构连线基本没有断裂，无子图独立出现，这是由各个学者间交流合作的特性导致的。说明作者之间或多或少存在一些交流与合作，但从图中可以看到几位高产作者如郑光、姜淼、郭洪涛的交流与联系较少。另一方面，该图左上角数据显式的节点数(N)为 499，连线数(E)为 293，密度(Density)为 0.0024。可见，我国文本挖掘的研究者们之间的交流合作意识仍然有待提高，学者们仍应多找机会进行学术交流与合作。

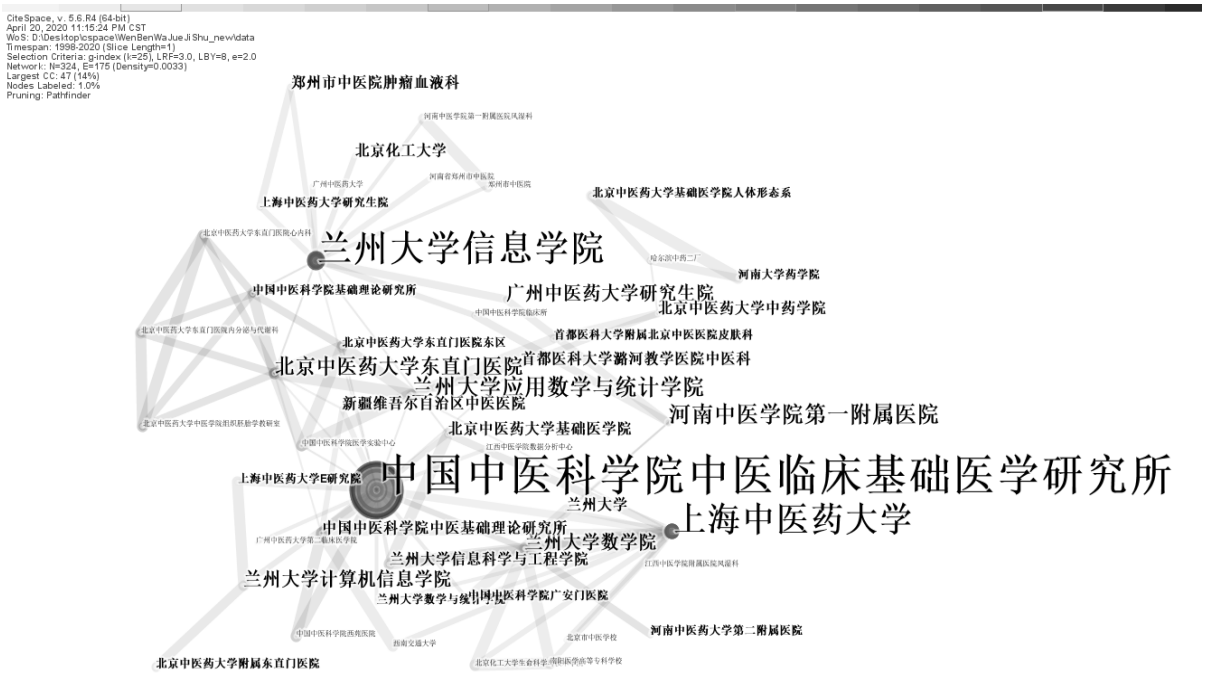

Figure 2. Linkage distribution of institutions
图 2. 机构共现可视化图谱

Figure 3. Linkage distribution of authors
图 3. 作者共现可视化图谱

(二) 研究主题与热点分析。某项研究的主题与热点是该研究方向和重点的集中体现，因此深层次挖掘研究的主题与热点信息对更加全面客观了解该研究内容具有十分显著的价值。而在某一研究中，关键词是一个十分有用的衡量一篇文献主要内容的标准。并且，关键词出现频率较高就可以反映该领域的研究热点[6]。此外，在使用关键词进行热点分析以聚类的方式进行时是以关键词共现分析作为基础，使用聚类统计学方法作为工具将关键词共现网络关系简化为聚类总数相对较少的过程[7]。而本文的研究侧重于通过该关键词聚类分析方法分析我国文本挖掘技术领域的研究热点，以可以客观分析出我国文本挖掘技术领域的研究热点。

在关键词聚类分析中，CiteSpace 的节点类型(Node Types)设置为关键词(Keyword)，其他参数保持不变。从而运行得到关键词关系可视化图谱，再对其选择 LLR 算法进行聚类，以得到图 4 所示的关键词聚类可视化图谱。此时从图左上角可以看到 Q = 0.6678 该值远大于 0.3 说明这个聚类效果是十分显著的；Mean Silhouette = 0.3873 也说明了该聚类的结构是合理的。并且，可以观察到其中呈现了“专利数据”“web 文本挖掘”“数据挖掘”“文本分类”“自然语言处理”“文本聚类”“hadoop”“技术主题”“中成药”“vsm 模型”“情感倾向分析”“信息抽取”“文本挖掘技术”“新闻”“隐结构模型”“深度学习”共 16 类，从而反映出了我国文本挖掘技术研究的热点并且该结果是合理显著的。

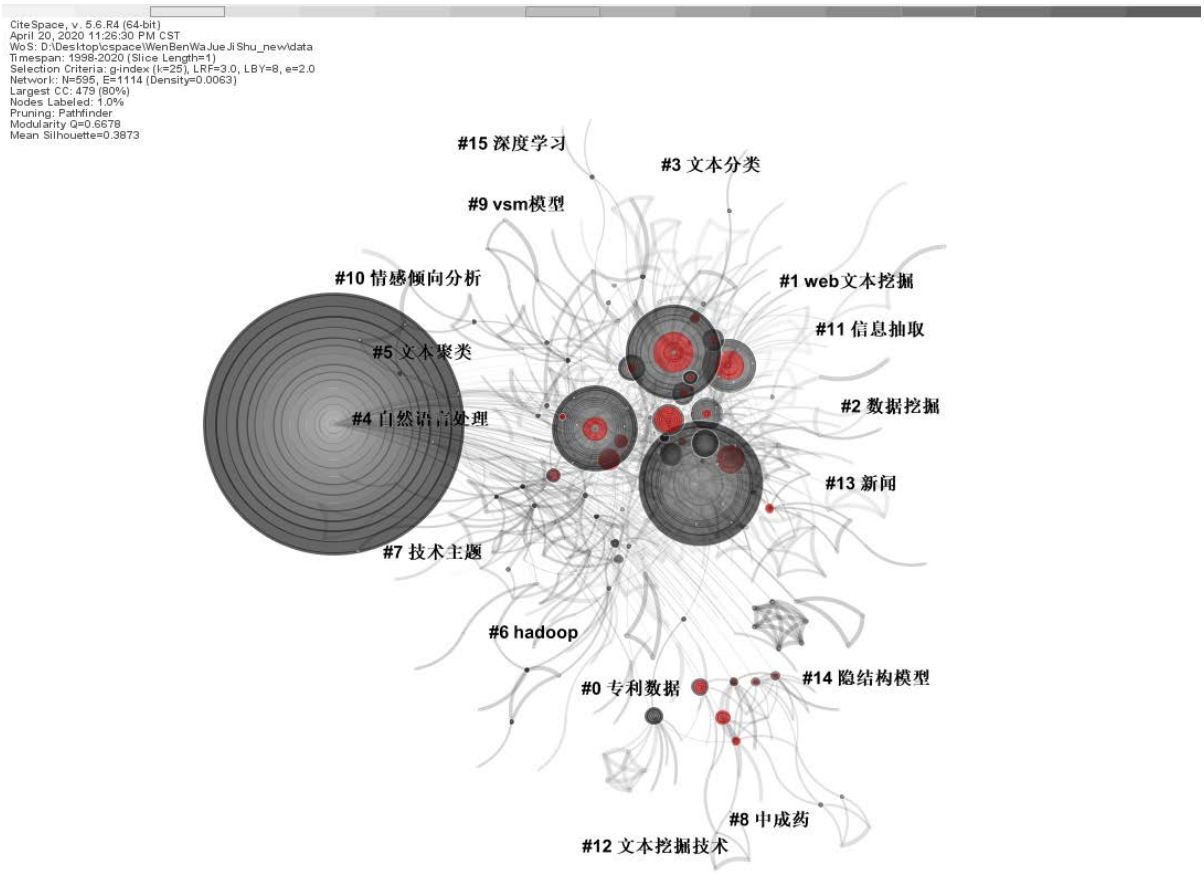

Figure 4. Keyword clustering
图 4. 关键词聚类可视化图谱

(三) 研究趋势分析。鉴于凸现出的关键词可以反映某个关键词在某段时间内被引用的次数突然激增，为此可以借助 CiteSpace 得到关键词凸现图以研究某一时间段内文本挖掘技术的研究趋势。如下图 5。图

片显示，在 2002~2012 年间的突现词为“web 挖掘”；2006~2010 年期间突现词为“文本分类”和“web 文本挖掘”；2011~212 年期间的突现词为“中成药”；2011~2013 年期间的突现词为“西药”；2012~2015 年期间的突现词为“数据分层算法”；2014~2020 年期间的突现词为“大数据”；2016~2020 年期间的突现词为“情感分析”。其中“大数据”和“情感分析”突现情况一直延续到现在，可以说明其为目前我国文本挖掘技术的主流发展趋势。

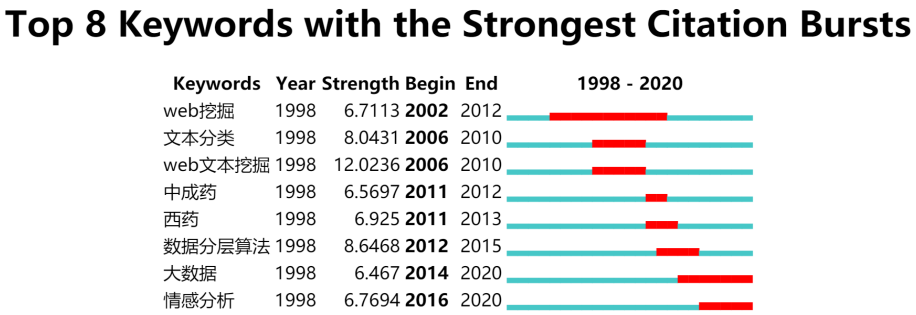

Figure 5. Keywords with the strongest citation bursts
图 5. 关键词凸现图

另一方面 CiteSpace 的强大功能给予了我们结合时序分析关键词的技术支持，而时序图即呈现该领域的研究关键词随时间变化的图示，其在一定程度上也可以反映各个时间段上的核心研究内容从而确定研究的大致趋势。如下图 6 所示，即为文本挖掘技术研究领域的关键字时序图。从图中不难观察到每个时间段内的重点关注对象不同，从而可以将我国文本挖掘技术的研究历程做一下划分。以更好的分析我国文本挖掘技术的研究发展。

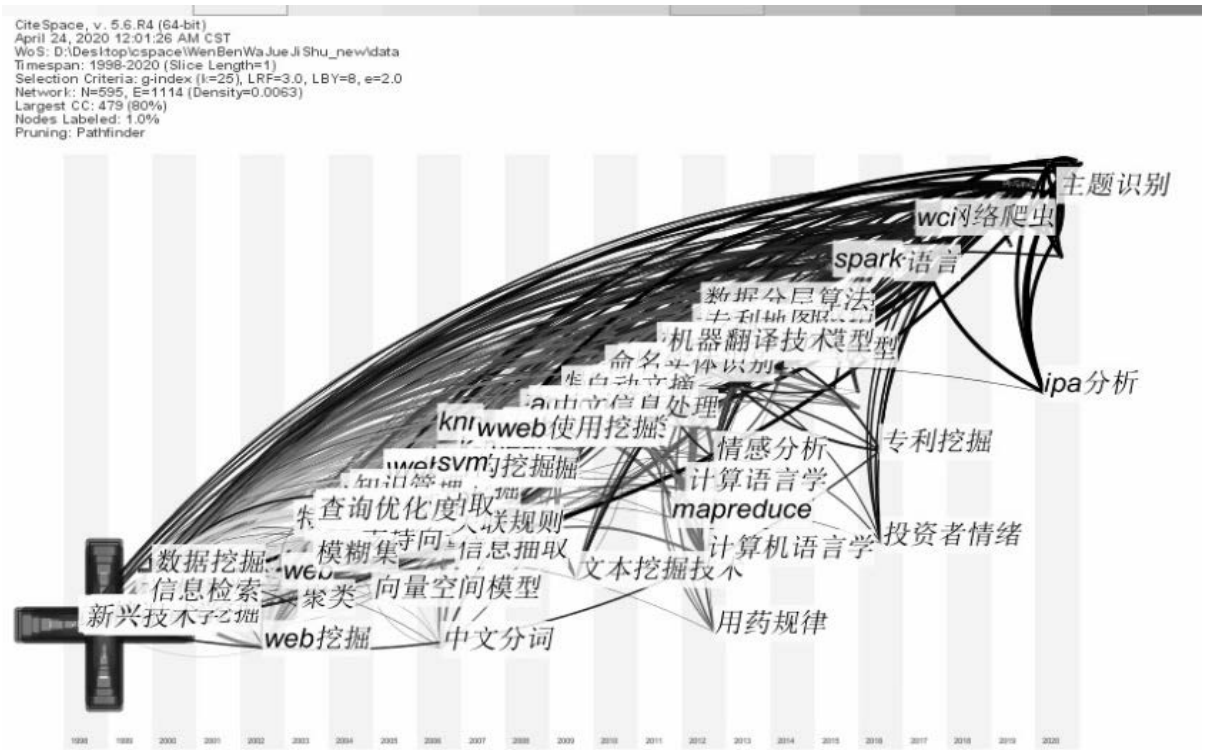

Figure 6. Keyword distribution over time
图 6. 关键词时序图

1) 1998~2002 年期间，可以看到文本挖掘研究主要的关键词有数据挖掘、信息挖掘、web 挖掘等，基于我国计算机技术的快速发展，文本挖掘获得了更优质的计算方式，促使文本挖掘技术的研究开始进入各位学者的视野。文本挖掘开始从数据信息检索向 web 端文本挖掘转移。

2) 2003 年之后。随着大数据技术的不断火热以及人工智能热的到来，而由于人工智能在中文领域的发展如语音识别、问答系统等都需要借助文本挖掘技术得到文本信息，再加之机器学习、深度学习技术的发展从而大大促进了我国文本挖掘这一基础技术提升，涌现出了大量的研究者。并且文本挖掘技术逐渐出现在各个领域，从之前的 web 文本挖掘逐渐到各个领域的文本挖掘技术如医药方面，专利挖掘方面、投资者情绪分析，再到机器翻译、数据分层算法、主题识别等的研究。从而大量文献的涌现也使得文本挖掘技术不断在更多的领域得到应用。

4. 总结及期望

对我国文本挖掘技术研究现状进行分析后发现：

第一，从文本挖掘研究的文献发文量角度看，整体发文量程上升趋势，2016 年及以前尤为显著，2017 年之后有所减少但整体数量仍保持在年 90 篇以上。说明文本挖掘技术自二十一世纪以来越来越受研究者青睐。

第二，从机构角度看，其分析结果表示研究文本挖掘的各机构间的联系比较少，主要以各自分散研发为主。而中国医科大学中医临床基础医学研究所和兰州大学信息学院以及上海中医大学对文本挖掘技术的研究贡献最为突出。为此应该加强各高产机构之间的交流合作，促进文本挖掘技术的进步。从发文作者角度看，经作者共被引分析发现，图普虽没呈现出完全独立的小团体，但集群间的联系十分稀少。文本挖掘技术研究做出显著贡献的有吕爱平、郭洪涛、正光和姜淼。以他们为中心的作者联系很紧密，而这些高产作者之间的联系依然很稀疏。从而，我国文本挖掘的研究者们之间的交流合作意识仍然有待提高，学者们仍应多找机会进行学术交流与合作。

第三，从研究热点及趋势的角度看，我国文本挖掘技术的研究热点在从最初的 web 文本挖掘，一步步走向各个领域的领域内文本挖掘，特别是医药领域。再到如今聚焦于大数据文本、情感分析、主题识别。这些热点也说明了文本挖掘目前正在大力发展的方向及研究热情主要聚焦的主题。因此，更有效地解决大数据下的文本挖掘、从文本中提取感情倾向以及提取文章主题中的问题仍然需要研究者投入热情，从而促进文本挖掘技术的进一步突破。促进依靠文本挖掘技术的人工智能等技术的发展。

基金项目

云南省高校数据化运营管理工程研究中心建设项目。

参考文献

[1] 周超峰. 文献计量常用软件比较研究[D]: [硕士学位论文]. 武汉: 华中师范大学, 2017.

[2] Chen, C., Song, I.Y., Yuan, X., et al. (2008) The Thematic and Citation Landscape of Data and Knowledge Engineering (1985-2007). *Data & Knowledge Engineering*, **67**, 234-259. https://doi.org/10.1016/j.datak.2008.05.004

[3] 李杰, 陈超美. CiteSpace 科技文本挖掘及可视化[M]. 北京: 首都经济贸易大学出版社, 2017: 1.

[4] Lippi, G. and Plebani, M. (2020) Procalcitonin in Patients with Severe Coronavirus Disease 2019 (COVID-19): A Meta-Analysis. *Clinica Chimica Acta*, **505**, 190-191. https://doi.org/10.1016/j.cca.2020.03.004

[5] 苗小燕, 张冲. 大中小学德育一体化研究的热点与发展趋势——基于 CNKI 数据库的 CITESPACE 分析[J]. 中国特殊教育, 2018(8): 85-90.

[6] 林德明, 陈超美, 刘则渊. 共被引网络中介中心性的 Zipf-Pareto 分布研究[J]. 情报学报, 2011, 30(1): 76-82.

[7] 钟伟金, 李佳, 杨兴菊. 共词分析法研究(三)——共词聚类分析法的原理与特点[J]. 情报杂志, 2008, 27(7): 118-120.